The Romanian Neuter Examined Through A Two-Gender N-Gram Classification System

Liviu P. Dinu[1, 3], Vlad Niculae[1, 3], Octavia-Maria Şulea[1, 2, 3]

Faculty of Mathematics and Computer Science[1], Faculty of Foreign Languages and Literatures[2]
Center for Computational Linguistics[3], University of Bucharest
ldinu@fmi.unibuc.ro, vlad@vene.ro, mary.octavia@gmail.com

Abstract

In this paper we look at the gender system of Romanian and investigate, using machine learning techniques, the validity of the traditional analysis according to which Romanian is a three gender language. We offer strong evidence in favor of the two gender system analysis proposed in (Bateman and Polinsky, 2010) with classification accuracy higher than the one previously obtained and diverge from the approaches found in the works of (Cucerzan and Yarowsky, 2003) and (Nastase and Popescu, 2009) on automated classification of Romanian nouns according to gender, which leads us to the best accuracy in discriminating the neuter.

Keywords: Computational Morphology, Romanian, Noun gender system

1. Introduction

Romanian has been traditionally seen as bearing three lexical genders: masculine, feminine and neuter, although it has always been known to have only two agreement patterns (for masculine and feminine). A recent analysis of the Romanian gender system described in (Bateman and Polinsky, 2010), based on older observations, argues that there are two lexically unspecified noun classes in the singular and two different ones in the plural and that what is generally called neuter in Romanian shares the class in the singular with masculines, and the class in the plural with feminines based not only on agreement features but also on form. Previous machine learning classifiers that have attempted to discriminate Romanian nouns according to gender have so far taken as input only the singular form, presupposing the traditional tripartite analysis. We propose a classifier based on two parallel support vector machines using n-gram features from the singular and from the plural which outperforms previous classifiers in its high ability to distinguish the neuter. The performance of our system suggests that the two-gender analysis of Romanian, on which it is based, is on the right track.

2. Gender and learnability in the case of Romanian

There have been different opinions with regard to how many genders there are in Romanian. The traditional analysis (Graur et al., 1966) envisions Romanian as the only Romance language bearing three genders (masculine, feminine, and neuter), whether the neuter was inherited from Latin, or (re)developped under the influence of Slavic languages (Rosetti, 1965; Petrucci, 1993). The three lexical genders are then mapped onto two agreement patterns, one for the singular, the other for the plural.

(Corbett, 1991) distinguished in Romanian three "controller genders" (into which nouns are divided), marked in the lexicon and corresponding to the three traditional genders, and two "target genders" (which are marked on adjectives, demonstratives, numerals, etc.), corresponding to the two agreement patterns for masculine and feminine, onto which the controller genders are mapped. (Farkas, 1990), on the other hand, describes the behavior of Romanian nouns, observing that masculines and feminines stay put, while neuters pattern with the masculines in the singular and with the feminines in the plural. Her analysis, although a three-gender account, lends itself to a two-gender interpretation of Romanian (Bateman and Polinsky, 2010, p. 51).

More recently, (Bateman and Polinsky, 2010) propose that Romanian has two noun classes in the singular and in the plural, that this categorization is not lexically specified, and that the division of nouns into classes in the singular is different from their division into classes in the plural. More preceisely, in the singular, masculine and neuter nouns are grouped together and separated from feminines, due to them being indistinguishable both in ending and in agreement pattern, while in the plural feminines and neuters are grouped together and separated from the masculines, due to the same reasons. The fact that what are considered neuter nouns in Romanian pattern with the masculines (in terms of agreement) in the singular and with the feminines in the plural has been a well know fact for Romanian linguists. What the analysis in (Bateman and Polinsky, 2010) puts forward is the idea that this is predictable through semantic and formal cues (singular endings) only, which enables the speaker to form the plural independent of the division of the nominal lexicon gender-wise.

In what follows we will investigate and attempt to validate the latter analysis employing machine learning techniques that render better results than previously obtained.

3. Approach

In order to automatically learn how to classify nouns of a particular language according to gender, one would need to first have an appropriate gender system analysis of that language. One of the two previous works on automatic classification of Romanian nouns according to gender, (Nastase

and Popescu, 2009), assumed the traditional analysis which envisions Romanian as having three distinct lexical genders and fitted machine learning methods only for the singular forms in an attempt to automatically learn how to identify the masculine, feminine, and neuter. The other, (Cucerzan and Yarowsky, 2003), also looked only at singular forms, but distinguished two classes instead of three: feminine vs. masculine and neuter.

In the light of analysis such as the one proposed in (Bateman and Polinsky, 2010), it is only natural for an automatic classification of singular Romanian noun forms to have a low performance in distinguishing neuters from masculines, as it should also be difficult to distinguish neuters from feminines in the plural form. We will, thus, look at singular and plural forms and investigate the hypothesis that neuters pattern with maculines in the singular and with feminines in the plural.

3.1. Dataset

The dataset we used is a Romanian language resource containing a total of 480,722 inflected forms of Romanian nouns and adjectives which was extracted from the text form of the morphological dictionary RoMorphDict (Barbu, 2008), where every entry has the following structure:

`form␣lemma␣description`

In the above, `form` refers to the fully inflected form of the noun and `description` refers to its morphosyntactic characterization. For the morphosyntactic description, the initial dataset uses the slash ('/') as a disjunct operator (or) meaning that 'm/n' stands for masculine or neuter, while the dash ('-') is used for the conjunct operator (and), with 'm-n' meaning masculine and neuter. In the Results section, we will see that some of the disjunct gender labels such as 'n/f' cause some problems in the extraction of the appropriate gender and subsequently in the automatic classifier system.

Since our interest was in the gender of nouns, we discarded all the adjectives listed and we isolated the nominative/accusative indefinite (without the enclitic article) form. We subsequently split them into singulars and plurals; the defective nouns were excluded. The entries which were labeled as masculine or feminine were used as training and validation data, while the neuters were left as the unlabeled test set.

The size of the training and validation data is 30,308 nouns, and the neuter test set consists of 9,822 nouns (each having a singular and a plural form).

3.2. Classifier and features

Our model consists of two binary linear support vector classifiers, one for the singular forms and another one for the plural forms. Each of these has a free parameter C that needs to be optimized to ensure good performance. Class labels are set to 0 for masculine and 1 for feminine nouns.

We extracted n-gram features from the masculine and feminine nouns, forming a large sparse matrix representation of the data. This ensures computational and memory efficiency when training the classifier. The feature extraction algorithm that builds this sparse matrix first iterates over all strings in the dataset, building a list of all the d n-grams that occur for every n between 1 and the maximum n-gram size. Then, in order to transform a noun from string form to such an n-gram representation, we simply turn it into a d-vector indicating how many times each feature occurs in the string. Alternatively, the vector can indicate simple binary occurrence instead of frequency. We also considered that the suffix might carry more importance so we added the '$' character at the end of each inflected form. This allows the downstream classifier to assign a different weight to the $(n-1)$-grams that overlap with the suffix.

Each possible combinations of parameters: n-gram length, use of binarization, addition of suffix, and the C regularization parameter from a small exponential space was evaluated using 10-fold cross-validation, for both singular and plural forms. The results are displayed in figure 1.

After the model has been selected and trained in this manner, the neuter nouns are plugged in and their singular forms are classified according to the singular classifier, while their plural forms are classified by the plural model.

The experiment was set up and run using the *scikit-learn* machine learning library for Python (Pedregosa et al., 2011). The implementation of linear support vector machines uses *liblinear* (Fan et al., 2008) behind the scenes, which scales to large numbers of samples and features. The n-gram extraction and grid search functionality is also available in *scikit-learn*.

4. Results

It is clear from the results of model selection in figure 1 that appending an artificial suffix character '$' improves the scores. In addition, it can be seen that in most cases, binarization does not help: we are simply better off keeping the counts. The maximum n-gram size of 5 seems a reasonable choice. When the size increases, it doesn't seem to help significantly, while for some parameter choices accuracy even begins to decrease at that point.

For 5-gram features, the SVM trained on the singular nouns, with the suffix character appended and without feature binarization, obtained an accuracy of 99.59%, with a precision score of 99.63%, a recall score of 99.80% and an F_1 score of 99.71%. The model trained on the masculine-feminine plural nouns under the same parameters scored an accuracy of 95.98%, with a precision score of 97.32%, a recall score of 97.05% and an F_1 score of 97.18%. We then moved on to check the classification results of the neuter forms, shown in figure 2.

Using the parameters agreed upon in the last paragraph, the results are that 99.17% of the neuters are classified as masculine. The plural neuter forms were classified as being feminine 93.22% of the time. While the binarized suffixless 3-gram plurals classifier fits our hypothesis better (96.56%), that model is not taken into consideration because it performs more poorly during validation, i.e. it is a poorer model of the masculine vs. feminine discrimination. We note that behaviour is much less erratic in the singular forms, where the best model for the discrimination also best fits our hypothesis. This encourages the idea that

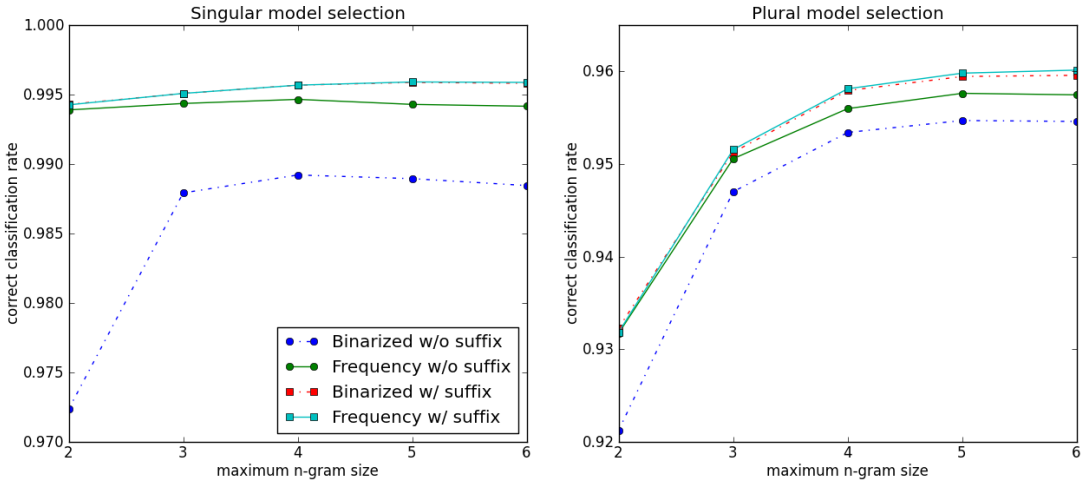

Figure 1: Model selection for all possible parameter choices. The y-label is the averaged correct classification rate estimated using 10-fold cross validation, showing only the best score for all C-values in the sampled interval.

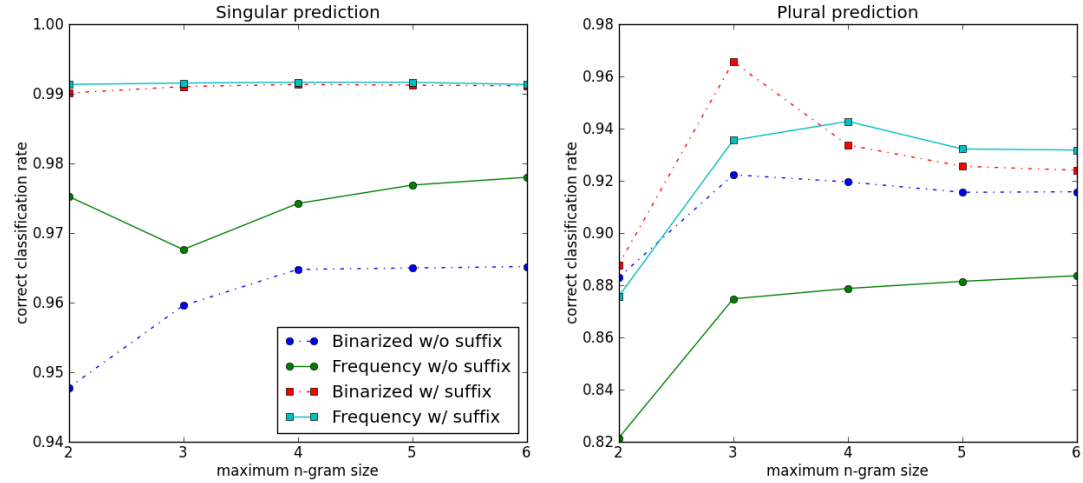

Figure 2: Results of applying the classifiers on the neuter forms. The y-label is the proportion of nouns classified as expected, i.e. as masculine in the left subplot, and as feminine in the right subplot.

s/p	f	m
m	9086	654
f	70	12

Table 1: Distribution of neuters as classified by the system. The upper right corner shows nouns classified as expected (masculine in the singular, feminine in the plural), while the lower right corner shows completely misclassified nouns (nouns that seem to be feminine in the singular and masculine in the plural). The other two fields appropriately show nouns misclassified in only one of the forms.

one should also examine the plural forms when studying the gender.

From the contingency table 1, we see that there are more misclassifications in the plural form of neuter nouns than in their singular form. In what follows, we will briefly analyze the misclassifications and see if there is any room for improvement or any blatant mistakes that can be rectified.

4.1. Analyzing Misclassifications

We first notice that 8 out of the 12 nouns that were completely misclassified are French borrowings which, although feminine in French, designate inanimate things. According to (Butiurca, 2005, p. 209), all feminine French nouns become neuter once they are borrowed into Romanian. The ones discussed here have the singular ending in 'é', written in Romanian without the accent, but retaining main stress as in French. Another of the 12, which also ends in an 'e' carrying main stress but not of French origin, is a noun formed from an acronym: *pefele* from *PFL*. There is also a noun (*coclaură–coclauri*) probably from the pre-Latin substratum, which is listed in Romanian dictionaries either as a pluralia tantum or as it is listed in the dataset. The final two are feminine singular forms wrongly labeled in the original corpus as being neuter or neuter/feminine. Looking at the entries in the original dataset for the last

two nouns (*levantin/levantină–levantinuri/levantine* and *bageac/bageacă–bageacuri/bageci*), we notice that the latter receives an 'n' tag for the singular form *bageacă*, which in (Collective, 2002) is listed as a feminine, and the former receives the 'n/f' tag, meaning *either a neuter, or a feminine* (Barbu, 2008, p. 1939), for both the neuter *levantin* and the feminine *levantină* singular form.

We further notice that, when the gender tag 'n/f' accompanies a singular form, a contradiction is stated. Seeing as Romanian has only two agreement patterns in the singular and plural (one for masculines and one for feminines) and that neuters agree like masculines in the singular and feminines in the plural, a feminine noun cannot be either neuter, and receive the masculine numeral *un* in the singular, or feminine, and receive the feminine numeral *o*. It can only be feminine. Through analoguous resoning, the tag 'n/m' accompanying a plural form is also absurd. By eliminating the second gender from the two disjunct labels of the RoMorphoDict lexicon when extracting the nouns for our classification experiments, we correctly tagged the neuter variants with 'n', but also wrongly tagged 5 feminine singular forms with 'n' and 7 masculine plural forms with 'n'. There are other misclassified nouns, from the other two groups, whose misclasification is due to an error in their initial gender label, for instance *algoritm–algoritmi* is shown to be a masculine in (Collective, 2002), however in the corpus it is tagged as neuter (together with the neuter variant *algoritm–algoritme*) and it subsequently appears to be misclassified in the plural as a masculine, which in fact it is.

Another problem causing the misclassification is represented by the hyphenated compound nouns, which are headed by the leftmost noun that also receives the number/gender inflection. Seeing as our classification system weighed more on the suffix, it was prone to fail in correctly clasifying them.

The final problem has to do with the 'uri' plural suffix which is par excellence a neuter plural desinence (Constantinescu-Dobridor, 2001, p. 44) in Romanian. There were 3044 neuter nouns ending in 'uri' in our testing set, out of which 2487 were correctly classified (as feminine in the plural). This means that out of the 654 neuter nouns misclassified in the plural (as masculine), 557 bore the suffix 'uri'. One reason for the confusion may be the fact that in our training set there were masculine nouns who seemed to receive the 'uri' suffix in the plural, but actually ended in 'ur' in the singular and received the (masculine) 'i' suffix in the plural (e.g. *balaur–balauri*).

5. Conclusions

Our results make a strong case for the analysis that the neuter in Romanian patterns with the masculine in the singular and with the feminine in the plural soley due to form and semantic content. Furthermore, our classification model outperforms the decision tree one described in the appendix of (Bateman and Polinsky, 2010) and the two classifiers of Romanian nouns according to gender previously constructed in terms of correctly distinguishing the neuter. This means that we have offered more than satisfactory reasons to consider the analysis of the Romanian gender system as a masculine-feminine one, with the Romanian neuter not being a proper gender, but a combination of the other two.

6. Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments. All authors contributed equally to this work. The research of Liviu P. Dinu was supported by the CNCS, IDEI - PCE project 311/2011, "The Structure and Interpretation of the Romanian Nominal Phrase in Discourse Representation Theory: the Determiners."

7. References

Ana-Maria Barbu. 2008. Romanian lexical databases: Inflected and syllabic forms dictionaries. In *Sixth International Language Resources and Evaluation (LREC'08)*.

Nicoleta Bateman and Maria Polinsky, 2010. *Romanian as a two-gender language*, chapter 3, pages 41–78. MIT Press, Cambridge, MA.

Doina Butiurca. 2005. Influenţa franceză. In *European Integration-Between Tradition and Modernity (EITM), Volume 1*, pages 206–212.

Collective. 2002. *Dicţionar ortografic al limbii române*. Editura Litera Internaţional.

Gheorghe Constantinescu-Dobridor. 2001. *Gramatica Limbii Române*. Editura Didactică şi Pedagogică Bucureşti.

Greville G. Corbett. 1991. *Gender*. Cambridge University Press.

S. Cucerzan and D. Yarowsky. 2003. Minimally supervised induction of grammatical gender. In *HLT-NAACL 2003*, pages 40–47.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, June.

Donka Farkas. 1990. Two cases of underspecification in morphology. *Linguistic Inquiry*, pages 539–550.

Alexander Graur, Mioara Avram, and Laura Vasiliu. 1966. *Gramatica Limbii Române*, volume 1. Academy of the Socialist Republic of Romania, 2nd edition.

Vivi Nastase and Marius Popescu. 2009. What's in a name? in some languages, grammatical gender. In *EMNLP*, pages 1368–1377. ACL.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, Oct.

Peter R. Petrucci. 1993. *Slavic features in the history of Romanian*. Ph.D. thesis.

Alexandru Rosetti. 1965. *Linguistica*. The Hague: Mouton.